ON THE EMERGENCE OF POSITION BIAS IN TRANSFORMERS

Xinyi Wu[1] **Yifei Wang**[2] **Stefanie Jegelka**[3,2] **Ali Jadbabaie**[1]
[1]MIT IDSS & LIDS [2]MIT CSAIL [3]TU Munich
{xinyiwu,yifei_w,stefje,jadbabai}@mit.edu

ABSTRACT

Recent studies have revealed various manifestations of position bias in transformer architectures, from the "lost-in-the-middle" phenomenon to attention sinks, yet a comprehensive theoretical understanding of how attention masks and positional encodings shape these biases remains elusive. This paper introduces a novel graph-theoretic framework to analyze position bias in multi-layer attention. Modeling attention masks as directed graphs, we quantify how tokens interact with contextual information based on their sequential positions. We uncover two key insights: First, causal masking inherently biases attention toward earlier positions, as tokens in deeper layers attend to increasingly more contextualized representations of earlier tokens. Second, we characterize the competing effects of the causal mask and relative positional encodings, such as the decay mask and rotary positional encoding (RoPE): while both mechanisms introduce distance-based decay within individual attention maps, their aggregate effect across multiple attention layers – coupled with the causal mask – leads to a trade-off between the long-term decay effects and the cumulative importance of early sequence positions. Through controlled numerical experiments, we not only validate our theoretical findings but also reproduce position biases observed in real-world LLMs. Our framework offers a principled foundation for understanding positional biases in transformers, shedding light on the complex interplay of attention mechanism components and guiding more informed architectural design.

1 INTRODUCTION

The attention mechanism is central to transformer architectures (Vaswani et al., 2017), which form the backbone of state-of-the-art foundation models, including large language models (LLMs). Its success lies in its ability to dynamically weigh input elements based on their relevance, enabling efficient handling of complex dependencies (Kim et al., 2017; Bahdanau et al., 2015). However, despite this widespread success, many questions remain unanswered regarding how these mechanisms process information and the artifacts they may introduce. Developing a deeper theoretical understanding of their inner workings is essential – not only to better interpret existing models but also to guide the design of more robust and powerful architectures.

One particularly intriguing aspect that demands such a theoretical investigation is *position bias*, i.e., the bias of the model to focus on certain regions of the input, which significantly impacts the performance and reliability of transformers and LLMs (Zheng et al., 2023; Wang et al., 2024; Hou et al., 2024). For instance, these models often suffer from the "lost-in-the-middle" problem, where retrieval accuracy significantly degrades for information positioned in the middle of the input sequence compared to information at the beginning or end (Liu et al., 2024; Zhang et al., 2024; Guo and Vosoughi, 2024). Similarly, in-context learning is highly sensitive to the order of illustrative examples: simply shuffling independently and identically distributed (i.i.d.) examples can lead to significant performance degradation (Min et al., 2022; Lu et al., 2022; Zhao et al., 2021). Moreover, recent research has also revealed that attention sinks (Xiao et al., 2024; Gu et al., 2025; Guo et al., 2024) – positions that attract disproportionately high attention weights – arise at certain positions regardless of semantic relevance, suggesting an inherent positional bias.

These empirical findings suggest that while transformers effectively encode and process positional information through the combined use of attention masks and positional encodings (PEs) (Wang et al., 2024; Fang et al., 2025), these design elements also appear to introduce systematic positional

biases, often independent of semantic content. This raises a fundamental and intriguing question about the role of positional information in attention mechanisms:

How do attention masks and positional encodings shape position bias in transformers?

To address the question, we propose a novel graph-theoretic framework for analyzing attention score distributions in multi-layer attention settings. Building upon Wu et al. (2024), we model attention masks as directed graphs, enabling rigorous mathematical analysis of attention patterns. This approach proves particularly powerful for studying multi-layer attention mechanisms, as it allows us to precisely quantify how each token's contextual representation is composed from information at different positions in the sequence. By tracking the information flow through the attention layers, we can systematically examine how positional biases emerge and propagate across layers, providing insights into the complex interplay between attention masks, PEs, and the network's depth.

Our contributions are summarized as follows:

- We show that causal masking in transformers inherently biases attention toward earlier positions in deep networks. This happens as tokens in deeper layers attend to increasingly more contextualized representations of earlier tokens, thereby amplifying the influence of initial positions.
- We uncover a nuanced interaction between causal masking and relative PEs, such as decay masks and rotary positional encoding (RoPE). Our findings highlight a trade-off in multi-layer attention networks, where local decay effects within individual layers are counterbalanced by the cumulative importance of early sequence positions. These results provide a deeper understanding of how PE and masking interact in deep attention-based architectures, with design implications about how to balance local and global context.
- We support our theoretical findings with experiments, empirically validating that deeper attention layers amplify the bias toward earlier parts of the sequence, while relative PEs partially mitigate this effect.

2 Problem Setup

Notation We use the shorthand $[n] := \{1, \ldots, n\}$. For a matrix M, we denote its i-th row by $M_{i,:}$ and its j-th column by $M_{:,j}$. Throughout the analysis in the paper, we formalize the attention mask to be a directed graph $\mathcal{G}$. Formally, we represent a directed graph with N nodes by $\mathcal{G}$ and let $E(\mathcal{G})$ be the set of directed edges of $\mathcal{G}$. A directed edge $(j, i) \in E(\mathcal{G})$ from node j to i in $\mathcal{G}$ means that in the attention mechanism, token j serves as a direct context for token i or token i attends to token j. The set $\mathcal{N}_i$ of all neighbors of node i is then $\{k : (k, i) \in E(\mathcal{G})\}$.

(Masked) attention mechanism Given the representation $X \in \mathbb{R}^{N \times d}$ of N tokens, the raw attention score matrix is computed as $Z = XW_Q(XW_K)^\top/\sqrt{d_{QK}}$, where $W_Q, W_K \in \mathbb{R}^{d \times d'}$ are the query and the key matrix, respectively, and $\sqrt{d_{QK}}$ is a temperature term to control the scale of raw attention scores. Without loss of generality, we assume $d_{QK} = 1$ in our analysis. To enforce a masked attention, we create a sparse attention matrix $A \in \mathbb{R}^{N \times N}$ based on Z whose sparsity pattern is specified by a directed graph $\mathcal{G}$: we normalize Z_{ij} among all allowed token interactions $(k, i) \in E(\mathcal{G})$ such that if $(j, i) \in E(\mathcal{G})$, $A_{ij} = \mathrm{softmax}_{\mathcal{G}}(Z_{ij}) = \frac{\exp(Z_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(Z_{ik})}$, and $A_{ij} = 0$ otherwise.

For our analysis, we consider single-head (masked) self-attention networks (SANs). The layerwise update rule can be written as

$$X^{(t+1)} = A^{(t)} X^{(t)} W_V^{(t)}, \tag{1}$$

where $A^{(t)} = \mathrm{softmax}_{\mathcal{G}^{(t)}}(X^{(t)} W_Q^{(t)} (X^{(t)} W_K^{(t)})^\top/\sqrt{d_{QK}})$ and $W_V^{(t)} \in \mathbb{R}^{d \times d'}$ is the value matrix. For simplicity, throughout the paper, we assume that $d = d'$ and $\mathcal{G}^{(t)} = \mathcal{G}$.

Relative Positional Encoding The *decay mask* represents the relative distance between two tokens by introducing an explicit bias favoring more recent tokens. Formally, it can be written as $D_{ij} = -(i - j)m$ if $j \leq i$ and 0 otherwise. Then applying the decay mask is essentially $A_{\text{decay}}^{(t)} = \mathrm{softmax}_{\mathcal{G}}(X^{(t)} W_Q^{(t)} (X^{(t)} W_K^{(t)})^\top + D)$. Note that while the decay mask formulation follows ALiBi (Press et al., 2022), it can be generalized to more complex variants such as KERPLE (Chi et al., 2022).

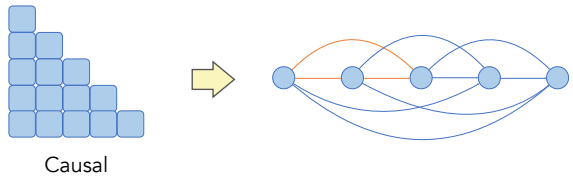

Figure 1: Causal graph and its corresponding directed graph used in the analysis (self-loops are omitted for clarity). A directed edge from token j to i indicates that i attends to j. The graph-theoretic formulation captures both direct and indirect contributions of tokens to the overall context, providing a comprehensive view of the token interactions under multi-layer attention.

Another way to encode the relative positional information is through *Rotary Positional Encoding (RoPE)* (Su et al., 2023), which applies a rotation to query and key embeddings by an angle proportional to the token's position index within the sequence. Formally, the rotation operation applied to each query or key $X_{i,:}W_{\{Q,K\}}$ can be written as $(\hat{X}_{\{Q,K\}})_{i,:} = X_{i,:}W_{\{Q,K\}}R^d_{\Theta,i}$, where $R^d_{\Theta,i}$ is the rotation matrix with a set of pre-defined base rotational angles $\Theta = \{0 \leq \theta_1 \leq \cdots \leq \theta_{d/2}\}$ (see Appendix A for the full form due to space limitations). Then the raw attention scores under RoPE Z_{RoPE} become $(Z_{\text{RoPE}})_{ij} = (X_{i,:}W_Q R^d_{\theta,i})(X_{j,:}W_K R^d_{\theta,j})^\top = X_{i,:}W_Q R^d_{\theta,i-j} W_K^\top X_{j,:}^\top$ which distorts the original raw attention scores based on the relative token distances. The final attention scores under RoPE are calculated as $A^{(t)}_{\text{RoPE}} = \text{softmax}_{\mathcal{G}}\left(Z^{(t)}_{\text{RoPE}}\right)$.

3 Main Results

In the transformer model, the attention mechanism is the sole module that allows tokens to interact with one another and incorporate contextual information from the sequence. It iteratively refines the contextual representation of each token across layers, allowing information to flow and accumulate based on relevance. This concept of contextualization through attention has its origins in the development of attention mechanisms, which predate transformers (Kim et al., 2017; Bahdanau et al., 2015). From the perspective of contextualization, the attention mechanism can be expressed in the following form (Kim et al., 2017):

$$X^{(t+1)}_{i,:} = \sum_{j=1}^{N} \underbrace{(A^{(t)} \cdots A^{(0)})_{ij}}_{\mathbb{P}^{(t)}(z_i = j | X^{(0)})} \cdot \underbrace{X^{(0)}_{j,:} W^{(0)}_V \cdots W^{(t)}_V}_{f^{(t)}(X^{(0)}_{z_i,:})}, \tag{2}$$

where z_i is a categorical latent variable with a sample space $\{1, \ldots, N\}$ that selects the input $X_{j,:}$ to provide context for token i. In this formulation, $A^{(t)}$ represents the attention matrix at layer t, $\mathbb{P}^{(t)}(z_i = j \mid X^{(0)})$ denotes the cumulative probability of selecting input token j as the context for token i at depth t , and $f^{(t)}(\cdot)$ is a learned transformation function.

This probabilistic formulation reveals two key aspects of the attention mechanism: it acts as both a context selector and a feature aggregator. As a selector, it assigns probabilities $\mathbb{P}^{(t)}$ that quantify the relevance of each token j to target token i at depth t. As an aggregator, it combines these selected contexts weighted by their respective probabilities $\mathbb{P}^{(t)}$ to form the contextualized representation $X^{(t)}$. Since position bias fundamentally manifests as systematic preferences in how tokens select and incorporate context from different positions, analyzing the attention mechanism's behavior is crucial for understanding these biases. By examining how attention masks and PEs affect the probability distribution $\mathbb{P}^{(t)}$, we can investigate how position-dependent patterns emerge and propagate through multi-layer attention in transformers.

Finally, we adopt the following assumptions in our analysis: **A1** There exists $C \in \mathbb{R}$ such that $\max_{t \in \mathbb{N}} \{\|W^{(t)}_Q\|_2, \|W^{(t)}_K\|_2\} \leq C$. **A2** The sequence $\{\| \prod_{t=0}^{k} W^{(t)}_V \|_2\}_{k=0}^{\infty}$ is bounded. In particular, **A1** assumes that the key and query weight matrices are bounded, which is crucial for efficient attention computation in practice (Alman and Song, 2023), whereas **A2** is to ensure boundedness of the node representations' trajectories $X^{(t)}$ for all $t \geq 0$ (Wu et al., 2024).

3.1 Attention Masks: A Graph-Theoretic View

We first analyze the case without PEs, focusing on the effect of attention masks. A graph-theoretic perspective offers a powerful framework for analyze multi-layer attention: the flow of attention across tokens can be represented as paths in a directed graph defined by the mask, where each path captures how information is transmitted between tokens (see Figure 1 for an illustration). The number of steps

in a path corresponds to the number of layers. By accounting for all such paths, we can quantify the cumulative influence of each token in the context computation of other tokens.

Through the graph-theoretic view, our first result states that for a causal mask $\mathcal{G}$, as tokens in deeper layers attend to increasingly more contextualized representations of earlier tokens, the context of each token converges exponentially toward the first token in the sequence.

Theorem 3.1. *Let $\mathcal{G}$ be the causal mask. Under* **A1**-**A2***, given $X^{(0)} \in \mathbb{R}^{N\times d}$, for every token $i \in [N]$, $\lim_{t\to\infty} \mathbb{P}^{(t)}(z_i = 1|X^{(0)}) = 1$. Moreover, there exist $0 < C, \epsilon < 1$ where $N\epsilon < 1$ such that $\mathbb{P}^{(t)}(z_i = j|X^{(0)}) \leq C(1-(j-1)\epsilon)^t$, for all $1 < j \leq i$ and $t \geq 0$.*

Theorem 3.1 reveals that in multi-layer causal attention, positional bias toward earlier sequence positions intensifies with depth – regardless of semantic content. This phenomenon arises from the nature of multi-layer attention: starting from the second layer, tokens no longer attend to raw inputs but instead to contextualized tokens, i.e., representations transformed by prior attention layers. Combined with the sequential structure of the causal mask, this iterative process amplifies the role of earlier tokens, as they influence later ones not only as direct context but also indirectly through intermediate tokens along the path.

3.2 Relative PEs: A Competing Decay Effect

Having analyzed how attention masks bias the model toward the beginning of the sequence, we now shift our focus to studying PEs, the other key mechanism for representing positional information in transformers. As the name suggests, relative PEs incorporate positional information by modifying the original attention scores in a way that reflects the relative positions of tokens. Among these, the decay mask (Press et al., 2022) explicitly introduces a distance-based decay effect into the attention mechanism. We begin by examining the effect of the decay mask on individual attention layers.

Lemma 3.2. *Consider the decay mask in Section 2 where $\mathcal{G}$ is causal. Under* **A1**-**A2***, given $X^{(0)} \in \mathbb{R}^{N\times d}$, there exists $C_{\max}, C_{\min} > 0$ such that $C_{\min}e^{-(i-j)m} \leq (A^{(t)}_{\text{decay}})_{ij} \leq C_{\max}e^{-(i-j)m}$, for all $j \leq i \in [N]$ and $t \geq 0$.*

Lemma 3.2 demonstrates that the decay mask introduces an exponential decay effect into each attention map, with the strength of the effect determined by the token distances. However, while this result characterizes the behavior of individual attention layers, the interaction between layers in a multi-layer setting leads to more intricate behaviors. Building on Lemma 3.2 , Theorem 3.3 examines the cumulative effect of the decay mask across multiple layers when combined with the causal mask.

Theorem 3.3. *Consider the decay mask in Section 2 where $\mathcal{G}$ is causal. Fix $T \geq 0$. Under* **A1**-**A2***, given $X^{(0)} \in \mathbb{R}^{N\times d}$, it holds for all $j \leq i \in [N]$ and $t \leq T$, $\mathbb{P}^{(t)}_{\text{decay}}(z_i = j|X^{(0)}) = \Theta\left(\binom{t+i-j}{i-j}e^{-(i-j)m}\right)$.*

Notably, if we denote $L(x) = \log(\binom{t+x}{x}e^{-xm})$, then $L(x)$ is not a monotone function of the distance x between two tokens. More precisely, under Stirling's approximation, the critical point, where the highest attention score occurs, is at $x^* = t/(e^m - 1)$. This means that increasing the decay strength m decreases x^*, making the model more biased towards recent tokens, whereas increasing the number of attention layers increases x^*, making the model more biased towards initial tokens.

Compared to Lemma 3.2, while the decay mask imposes a stronger decay effect on earlier tokens within individual attention layers, these tokens gain more cumulative importance across multiple layers. This trade-off between layer-wise decay and cross-layer accumulation transforms the initially monotonic decay pattern within each attention map into a more intricate, non-monotonic behavior when aggregated throughout the network.

We now turn our attention to another popular form of relative positional encoding: RoPE (Su et al., 2023). While RoPE's inherent complexity has made a clear theoretical understanding challenging, recent empirical observations in Barbero et al. (2024b) suggest that in practice, LLMs tend to predominantly utilize feature dimensions that rotate slowly. This phenomenon introduces additional structure, enabling a more refined analysis of RoPE's effects by focusing on these slowly rotating feature dimensions. For simplicity and without loss of generality, we consider the case where only the slowest-rotating feature dimensions with base rotational angle θ_1 are used by the model, i.e. effectively reducing the embedding dimension to $d = 2$. See Appendix H for more results on the general case $d \geq 2$.

Recall from Section 2 that RoPE operates by rotating the original query and key embeddings by an angle proportional to the token's position index within the sequence. Similar to the decay mask, which incorporates distance-based decay into attention scores, RoPE adjusts raw attention scores via these rotations. To formalize this relationship mathematically, we define the original angle between query $q_i^{(t)} := X_{i,:}^{(t)} W_Q^{(t)}$ and key $k_j^{(t)} := X_{j,:}^{(t)} W_K^{(t)}$ as $\phi_{i,j}^{(t)}$. Then the following result analyzes how RoPE's position-dependent rotations systematically modify the computation of attention scores.

Lemma 3.4. *Let $\mathcal{G}$ be the causal mask and $d = 2$. Suppose for $t \geq 0$, $\|q_i^{(t)}\|_2, \|k_j^{(t)}\|_2 > 0$, and $|\phi_{i,j}^{(t)}| \leq \delta\theta_1$, where $\delta > 0$ and $(\delta + N - 1)\theta_1 \leq \pi$. Then under* **A1**-**A2**, *given $X^{(0)} \in \mathbb{R}^{N \times d}$, there exists $C_{\max}, C_{\min}, c, c' > 0$ such that for all $j \leq i \in [N]$, $C_{\min} e^{-c(i-j)^2\theta_1^2} \leq (A_{\mathrm{RoPE}}^{(t)})_{ij} \leq C_{\max} e^{-c'(i-j)^2\theta_1^2}$.*

The result shows that by solely leveraging feature dimensions that rotate slowly, RoPE effectively induces a distance-based decay effect, which aligns with the intuition in Su et al. (2023). It is also worth noting that the decay effect induced by RoPE is significantly smaller compared to that of the decay mask (Lemma 3.2), as the base rotational angle θ_1 is typically chosen to be small, i.e. $\approx 1/10000$ per token (Su et al., 2023; Dubey and et al., 2024), resulting in a more gradual decay.

However, similar to the case of the decay mask, when considering the effect of RoPE across multiple layers of attention, the long-term decay effects within individual layers are counteracted by the increasing influence of earlier tokens given by the causal mask.

Theorem 3.5. *Fix $T > 0$. Under the same conditions as in Lemma 3.4 for $t \leq T$, given $X^{(0)} \in \mathbb{R}^{N \times d}$, there exists $c > 0$ such that for all $j \leq i \in [N]$ and $t \leq T$, $\mathbb{P}_{\mathrm{RoPE}}^{(t)}(z_i = j | X^{(0)}) = \Theta\left(\binom{t+i-j}{i-j} e^{-c(i-j)^2\theta_1^2}\right)$.*

Again, if we write $L(x) = \log(\binom{t+x}{x} e^{-x^2\theta_1^2})$, then, by implicit differentiation, the critical point x^* is an increasing function of the depth t and a decreasing function of the base rotational angle θ_1 (see Appendix G). This implies that increasing the base rotational angle θ_1 reduces the optimal distance x^*, amplifying the long-term decay effect and causing tokens to focus more on nearby tokens. In contrast, increasing the number of attention layers t increases x^* and hence deeper models become more biased toward initial tokens.

4 Experiments

In this section, we validate our theoretical findings via carefully designed numerical experiments. To ensure a controlled setup that enables precise manipulation of positional biases in the data, we adopt the synthetic data-generating process and simplified self-attention network framework proposed in Reddy (2024). This setup allows us to systematically isolate and examine the effects of different components on the emergence of position bias.

Task structure Following Reddy (2024), we adopt the following information retrieval task: The model is trained to predict the label y_{query} of a target x_{query} using the cross-entropy loss, given an alternating sequence of n items and n labels: $x_1, y_1, \ldots, x_n, y_n, x_{\text{query}}$. The sequence is embedded in d dimensions. Each x_i is sampled from a Gaussian mixture model with K classes, and y_i is the corresponding class label assigned prior to training from the total L labels ($L \leq K$). The burstiness B is the number of occurrences of x_i from a particular class in an input sequence. Importantly, at least one item in the context belongs to the same class as the query. To control position bias in the training data, x_{query} can either be explicitly assigned to the class of a specific x_i, introducing position-dependent bias in the data, or randomly assigned to the class of any x_i, simulating a scenario without position bias in the data.

Tracking position bias To quantify position bias, we evaluate model performance using sequences containing novel classes not seen during training. Specifically, by generating new class centers for the Gaussian mixture and randomly assigning one of the L existing labels to these novel classes, we ensure that the model relies on contextual information rather than memorized class features. Crucially, we can systematically vary the position of the correct answer within test sequences to measure retrieval accuracy changes, thereby isolating and quantifying position-dependent biases in the model's behavior.

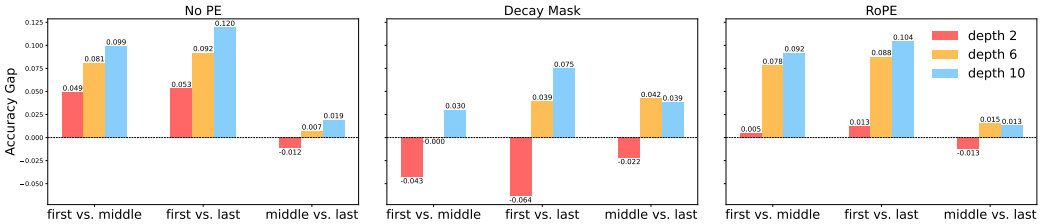

Figure 2: Position bias arising solely from the architectural design of the attention mechanism, with **no positional bias in the training data**. a vs. b denotes the gap for the case $[\boldsymbol{a}, b] - [\boldsymbol{b}, a]$, where bar magnitude indicates gap size, positive indicates bias toward earlier position, and negative indicates bias toward later position. Deeper attention amplifies the bias toward earlier tokens, regardless of the PE used. Furthermore, decay mask introduce stronger distance-based decay effects that increase focus on recent tokens than RoPE.

Network architecture The input sequences are passed through an attention-only network followed by a classifier. Each attention layer has one attention head. The classifier is then a three-layer MLP with ReLU activations and a softmax layer which predicts the probabilities of the L labels.

Following Reddy (2024), we set $n = 8$ and $d = 64$. Additional experimental details are provided in Appendix I. Despite our use of a simplified experimental setup, we observe the emergence of key phenomena documented in real-world LLMs, such as the "lost-in-the-middle" phenomenon (Appendix J.1) and the formation of attention sinks (Appendix J.2). This convergence between our controlled environment and real-world observations validates our choice of abstraction, suggesting that we have preserved the essential mechanisms driving position bias while enabling systematic investigation.

4.1 The Effects of Depth and Relative PEs

To investigate the position bias arising solely from the architectural design of the attention mechanism, we use training sequences without positional bias, where the position of x_i sharing the same class as x_{query} is uniformly random in $\{1, 2, \ldots, n\}$. To evaluate the position bias in the trained model, we construct test sequences of the form $[\boldsymbol{a}, b]$. Here, the bolded term $\boldsymbol{a}$ explicitly marks the correct position, ensuring y_a matches y_{query}, while position b serves as a baseline. In these sequences, x_a and x_b are identical vectors, allowing us to control for the influence of semantic information on the model's retrieval accuracy. We then measure the retrieval accuracy gap between pairs of sequences where the content at positions a and b is identical, but the correct position varies. This gap, defined as $[\boldsymbol{a}, b] - [\boldsymbol{b}, a]$, quantifies the model's positional preference independent of semantic information. To perform this evaluation, we construct three pairs of test sets, each containing $10,000$ sequences: [**first**, middle] vs. [**middle**, first], [**first**, last] vs. [**last**, first], and [**middle**, last] vs. [**last**, middle]. Here "first" (position 1), "middle" (position $n/2$), and "last" (position n) denote fixed positions within a sequence.

Figure 2 shows the average results over five runs, where a vs. b denotes the gap $[\boldsymbol{a}, b] - [\boldsymbol{b}, a]$. The magnitude of each bar represents the size of the performance gap, and the sign of each bar reflects the direction of the bias: a positive sign indicates a bias toward earlier positions, while a negative sign indicates a bias toward later positions. We highlight several key observations. First, increasing model depth consistently amplifies the bias toward earlier parts of the sequence, regardless of the PE used. Also note that the performance gap between the middle and last positions is notably smaller than that involving the first position. This aligns with our theory, which suggests that as the model focuses more on the initial part of the sequence, information near the sequence's end becomes less distinguishable, consistent with the patterns observed in Barbero et al. (2024a). Furthermore, both the decay mask and RoPE introduce distance-based decay effects that reduce the bias toward the beginning induced by the causal mask and increase the focus on recent tokens. However, the decay effect induced by the decay mask is substantially more pronounced than that by RoPE, as predicted by our theory.

5 Conclusion

In this paper, we study position bias in transformers through a probabilistic and graph-theoretic lens, developing a theoretical framework that quantifies how positional information influences context construction across multi-layer attention. Our analysis reveals two key findings about position bias in transformers: the causal mask's inherent bias toward earlier tokens, as deeper layers increasingly attend to these positions through iterative attention, and the interplay between causal masking and relative positional encodings, which results in a nuanced, non-monotonic balance between

distance-based decay effects and the cumulative influence of earlier positions. By deepening our understanding of how architectural choices in transformers shape positional dependencies, our work provides a foundation for designing attention mechanisms with predictable and task-aligned positional properties.

REFERENCES

Josh Alman and Zhao Song. Fast attention requires bounded entries. In *NeurIPS*, 2023.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

Federico Barbero, Andrea Banino, Steven Kapturowski, Dharshan Kumaran, Joao G.M. Ara'ujo, Alex Vitvitskyi, Razvan Pascanu, and Petar Velivckovi'c. Transformers need glasses! information over-squashing in language tasks. In *NeurIPS*, 2024a.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Velivckovi'c. Round and round we go! what makes rotary positional encodings useful? *ArXiv*, abs/2410.06205, 2024b.

Ta-Chung Chi, Ting-Han Fan, Peter J. Ramadge, and Alexander I. Rudnicky. Kerple: Kernelized relative positional embedding for length extrapolation. In *NeurIPS*, 2022.

Abhimanyu Dubey and et al. The llama 3 herd of models. *ArXiv*, abs/2407.21783, 2024.

Lizhe Fang, Khashayar Gatmiry Yifei Wang, Lei Fang, and Yisen Wang. Rethinking invariance in in-context learning. In *ICLR*, 2025.

Murray Glanzer and Anita R. Cunitz. Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, 1966.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *ICLR*, 2025.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I. Jordan, and Song Mei. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in llms. 2024.

Xiaobo Guo and Soroush Vosoughi. Serial position effects of large language models. *ArXiv*, 2024.

Michael A.K. Halliday. *An Introduction to Functional Grammar*. 2004.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena A. Jäger, and Lisa Beinborn. Multilingual language models predict human reading behavior. In *NAACL*, 2021.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. In *ECIR*, 2024.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. Structured attention networks. In *ICLR*, 2017.

Yuxuan Li, Jesse K. Pazdera, and Michael J. Kahana. Eeg decoders track memory dynamics. *Nature Communications*, 2024.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *ACL*, 2022.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP*, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *ICLR*, 2022.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *ICLR*, 2024.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. 2023.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach. *ArXiv*, abs/2407.01100, 2024.

Xinyi Wu, Amir Ajorlou, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the role of attention masks and layernorm in transformers. In *NeurIPS*, 2024.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *ICLR*, 2024.

Zhenyu (Allen) Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. *ArXiv*, abs/2403.04797, 2024.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *ICML*, 2021.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, 2023.

A Rotational matrix used for RoPE

$$R^d_{\Theta,i} = \begin{bmatrix} \cos i\theta_1 & \sin i\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ -\sin i\theta_1 & \cos i\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos i\theta_2 & \sin i\theta_2 & \cdots & 0 & 0 \\ 0 & 0 & -\sin i\theta_2 & \cos i\theta_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & & 0 & 0 & \cos i\theta_{d/2} & \sin i\theta_{d/2} \\ 0 & 0 & & 0 & 0 & -\sin i\theta_{d/2} & \cos i\theta_{d/2} \end{bmatrix}$$

B Proof of Theorem 3.1

B.1 Auxiliary results

Lemma B.1. *Under* **A1**-**A2**, *there exists* $\epsilon > 0$ *such that* $A^{(t)}_{ij} \geq \epsilon$ *for all* $t \geq 0$, $(j, i) \in E$.

Proof. Writing (1) recursively, we get that the token trajectories

$$X^{(t+1)} = A^{(t)} ... A^{(0)} X^{(0)} W^{(0)}_V ... W^{(t)}_V \,, \tag{3}$$

stay uniformly bounded for all $t \geq 0$ by **A2**. Then it follows from **A1** that there exists $C \in \mathbb{R}$ such that for all $t \geq 0$,

$$\begin{aligned} \left\| \left(X^{(t)} W_Q^{(t)} \right)_{i,:} \right\|_2 &= \left\| X_{i,:}^{(t)} W_Q^{(t)} \right\|_2 \leq C \,, \\ \left\| \left(X^{(t)} W_K^{(t)} \right)_{i,:} \right\|_2 &= \left\| X_{i,:}^{(t)} W_K^{(t)} \right\|_2 \leq C \,. \end{aligned} \tag{4}$$

Hence for all $i, j \in [N]$,

$$-C^2 \leq (X^{(t)} W_Q^{(t)} (X^{(t)} W_K^{(t)})^\top)_{ij} \leq C^2 \,.$$

This implies that there exists $\epsilon > 0$ such that $A_{ij}^{(t)} \geq \epsilon$ for all $(j, i) \in E$. □

B.2 Proof of Theorem 3.1

We denote $P^{(t)} := A^{(t)} \cdots A^{(0)}$. It suffices to show that there exists $0 < C < 1$ and $0 < \epsilon < 1$ such that

$$P_{ij}^{(t)} \leq C(1 - (j-1)\epsilon)^t \tag{5}$$

for all $1 < j \leq i$ and $t \geq 0$.

The proof will go by induction:

Base case By Lemma B.1, it follows that

$$P_{ij}^{(0)} \leq (1 - \epsilon)$$

for all $1 < j \leq i$. Then let $C := 1 - \epsilon$.

Induction step Assume that (5) holds, it follows that for all $1 < j \leq i$.

$$P_{ij}^{(t+1)} = \sum_{k=j}^{i} A_{ik}^{(t)} P_{kj}^{(t)} \leq (1 - (j-1)\epsilon) C (1 - (j-1)\epsilon)^t = C(1 - (j-1)\epsilon)^{t+1} \,.$$

From above, we conclude the theorem.

C Proof of Lemma 3.2

Fix $t \geq 0$. Let

$$Z_{ij}^{(t)} = (X^{(t)} W_Q^{(t)})_{i,:} (X^{(t)} W_K^{(t)})_{:,j} \,.$$

Following from Lemma B.1, there exists $I_{\min}, I_{\max} \in \mathbb{R}$ such that for all $j \leq i \in [N]$,

$$Z_{ij}^{(t)} \in [I_{\min}, I_{\max}] \,.$$

Consider the denominator in the $\mathrm{softmax}(\cdot)$ operation in the calculation of $(A_{\mathrm{decay}}^{(t)})_{ij}$:

$$\begin{aligned} \sum_{k=1}^{i} e^{Z_{ik}^{(t)} - (i-k)m} &\geq e^{I_{\min}} \sum_{k=0}^{i} e^{-(i-k)m} \\ &= e^{I_{\min}} \frac{1 - e^{-(i+1)m}}{1 - e^{-m}} \\ &\geq e^{I_{\min}} \frac{1 - e^{-2m}}{1 - e^{-m}} \\ &= e^{I_{\min}} (1 + e^{-m}) \end{aligned}$$

and

$$\sum_{k=1}^{i} e^{Z_{ik}^{(t)}-(i-k)m} \leq e^{I_{\max}} \sum_{k=0}^{\infty} e^{-km}$$
$$= \frac{e^{I_{\max}}}{1-e^{-m}}$$

It follows that

$$(A_{\text{decay}}^{(t)})_{ij} \leq \frac{e^{I_{\max}-(i-j)m}}{e^{I_{\min}}(1+e^{-m})} = C_{\max} e^{-(i-j)m}$$

and

$$(A_{\text{decay}}^{(t)})_{ij} \geq \frac{e^{I_{\min}-(i-j)m}}{e^{I_{\max}}/(1-e^{-m})} = C_{\min} e^{-(i-j)m}$$

where $C_{\max} := e^{(I_{\max}-I_{\min})}/(1+e^{-m})$ and $C_{\min} := (1-e^{-m})e^{(I_{\min}-I_{\max})}$.

D Proof of Theorem 3.3

Note that in the causal graph $\mathcal{G}$, there are $\binom{t+i-j}{i-j}$ paths of length $t+1$ from token j to token i.

Since going from token j to token i in the causal graph, the connectivity patterns ensure that the token indices along the path are non-decreasing, i.e. if we denote the directed path as $(j, l_1), (l_1, l_2), ..., (l_t, i)$, it holds that $j \leq l_1 \leq l_2 \leq ... \leq l_t \leq i$. Together with Lemma 3.2, we conclude the theorem statement.

E Proof of Lemma 3.4

Fix $t \geq 0$. Denote the angle after rotation to be $\psi_{i,j}^{(t)}$. Then it follows from the definition of RoPE that

$$\psi_{i,j}^{(t)} = \phi_{i,j}^{(t)} - (i-j)\theta_1 \,.$$

Thus

$$|\psi_{i,j}^{(t)}| = |\phi_{i,j}^{(t)} - (i-j)\theta_1| \geq ||(i-j)\theta_1| - |\phi_{i,j}^{(t)}|| \geq |(i-j) - \delta|\theta_1 \,.$$

$$|\psi_{i,j}^{(t)}| = |\phi_{i,j}^{(t)} - (i-j)\theta_1| \leq |(i-j)\theta_1| + |\phi_{i,j}^{(t)}| \leq (i-j+\delta)\theta_1 \,.$$

Let the original query i and key j embeddings be $q_i^{(t)} := X_{i,:}^{(t)} W_Q^{(t)}$ and $k_j^{(t)} := X_{j,:}^{(t)} W_K^{(t)}$, respectively, and the corresponding query i and key j embeddings after rotation be $q_i'^{(t)}$ and $k_j'^{(t)}$, respectively.

Since $\langle q_i'^{(t)}, k_j'^{(t)} \rangle = \|q_i^{(t)}\|_2 \|k_j^{(t)}\|_2 \cos \psi_{i,j}^{(t)}$, it follows from that there exists $C_{\min}, C_{\max} \geq 0$ such that for all $i, j \in [N]$,

$$C_{\min} \cos((i-j+\delta)\theta_1) \leq \langle q_i'^{(t)}, k_j'^{(t)} \rangle \leq C_{\max} \cos(|(i-j)-\delta|\theta_1) \,.$$

Since for $|x| \leq \pi$ there exists $c > 0$ such that $1 - x^2/2 \leq \cos x \leq 1 - x^2/c$, we get that

$$C_{\min} \left(1 - \frac{((i-j)+\delta)^2\theta_1^2}{2}\right) \leq \langle q_i'^{(t)}, k_j'^{(t)} \rangle \leq C_{\max} \left(1 - \frac{((i-j)-\delta)^2\theta_1^2}{c}\right) \,.$$

and hence

$$C_{\min} \left(1 - \delta^2\theta_1^2 - (i-j)^2\theta_1^2\right) \leq \langle q_i'^{(t)}, k_j'^{(t)} \rangle \leq C_{\max} \left(1 - \frac{((i-j)^2/2 - \delta^2)\theta_1^2}{c}\right) \,.$$

Consider $Y_i^{(t)} = \sum_{k=1}^{i} e^{\langle q_i'^{(t)}, k_j'^{(t)} \rangle}$. Then by (4), we get that there exists $Y_{\max}, Y_{\min} > 0$ such that

$$Y_{\max} \leq Y_i^{(t)} \leq Y_{\min} \,.$$

We thus conclude the statement.

F PROOF OF THEOREM 3.5

Notice that

$$(P_{\mathrm{RoPE}}^{(t)})_{ij} = \sum_{l_1 \le \cdots \le l_{t-1} \in [N]^{t-1}} A_{il_{t-1}}^{(t-1)} A_{l_{t-1}l_{t-2}}^{(t-2)} \cdots A_{l_1 j}^{(0)} \tag{6}$$

Given that when $\mathcal{G}$ is the causal graph, due to the connectivity the directed path of length t from token j to token i must be non-decreasing, i.e. $j \le l_1 \le l_2 \le \cdots \le l_{t-1} \le i$, and there would be in total $\binom{t+i-j}{i-j}$ such paths. For each such path $j \le l_1 \le l_2 \le \cdots \le l_{t-1} \le i$, notice that by Lemma 3.4, we get that fix $T \ge 0$, there exists $C_{\min}, C_{\max} > 0$ such that

$$A_{i,l_{t-1}}^{(t-1)} A_{l_{t-1},l_{t-2}}^{(t-2)} \cdots A_{l_1,j}^{(0)} \ge C_{\min} e^{-c((i-l_{t-1})^2 + (l_{t-1}-l_{t-2})^2 + \cdots + (l_1-j)^2)\theta_1^2} \tag{7}$$

and

$$A_{i,l_{t-1}}^{(t-1)} A_{l_{t-1},l_{t-2}}^{(t-2)} \cdots A_{l_1,j}^{(0)} \le C_{\max} e^{-c'((i-l_{t-1})^2 + (l_{t-1}-l_{t-2})^2 + \cdots + (l_1-j)^2)\theta_1^2} \tag{8}$$

From (7), since $j \le l_1 \le l_2 \le \cdots \le l_{t-1} \le i$, we further get that

$$A_{i,l_{t-1}}^{(t-1)} A_{l_{t-1},l_{t-2}}^{(t-2)} \cdots A_{l_1,j}^{(0)} \ge C_{\min} e^{-c(i-j)^2\theta_1^2}\,, \tag{9}$$

and similarly

$$A_{i,l_{t-1}}^{(t-1)} A_{l_{t-1},l_{t-2}}^{(t-2)} \cdots A_{l_1,j}^{(0)} \le C_{\max} e^{-\frac{c'}{2}(i-j)^2\theta_1^2}\,. \tag{10}$$

G IMPLICIT DIFFERENTIATION OF x WITH RESPECT TO θ_1 AND t

Recall that under RoPE,

$$L(x) = \log\left(\binom{t+x}{x} e^{-x^2\theta_1^2}\right).$$

Then by Stirling's approximation,

$$L(x) \approx ((t+x)\log(t+x) - (t+x)) - (x\log x - x) - \theta_1^2 x^2\,,$$

and thus

$$L'(x) = \log\left(\frac{t+x}{x}\right) - 2\theta_1 x\,.$$

Taking implicit differentiation of t:

$$\frac{\partial}{\partial t}\log\left(\frac{t+x}{x}\right) = \frac{-t}{x(x+t)}\frac{\partial x}{\partial t} + \frac{1}{x+t}$$

and

$$\frac{\partial}{\partial t} 2\theta_1 x = 2\theta_1 \frac{\partial x}{\partial t}.$$

So let

$$\frac{1}{x+t} = (2\theta_1 + \frac{t}{(x+t)x})\frac{\partial x}{\partial t}$$

and thus

$$\frac{\partial x}{\partial t} = \frac{1}{2\theta_1(x+t) + \frac{t}{x}} > 0\,.$$

Taking implicit differentiation of θ_1:

$$\frac{\partial}{\partial \theta_1}\log\left(\frac{t+x}{x}\right) = \frac{-t}{x(x+t)}\frac{\partial x}{\partial \theta_1}$$

and

$$\frac{\partial}{\partial \theta_1} 2\theta_1 x = 2x + 2\frac{\partial x}{\partial \theta_1}\theta_1.$$

So let

$$\left(2\theta_1 + \frac{t}{x(x+t)}\right)\frac{\partial x}{\partial \theta_1} = -2x\,,$$

and thus

$$\frac{\partial x}{\partial \theta_1} = \frac{-2x}{2\theta_1 + \frac{t}{x(x+t)}} < 0\,.$$

Hence we observe that x^* is an increasing function of t and a decreasing function of θ_1. This implies that increasing the base rotational angle θ_1 reduces the optimal distance x^*, amplifying the long-term decay effect and causing tokens to focus more on nearby tokens. In contrast, increasing the number of attention layers t increases x^* and hence deeper models become more biased toward initial tokens.

H THE EFFECT OF ROPE: CASE FOR $d \geq 2$

In this section, we present a generalized version of Theorem 3.5 for the case $d \geq 2$.

Let the query q and key k be vectors in $\mathbb{R}^d$, where d is even, and let ϕ be the angle between q and k, which we assume to be well-defined, with:

$$\cos\phi = \frac{\langle q, k\rangle}{\|q\|_2\|k\|_2}\,.$$

Define the length-2 segments of query q and and key k as

$$q_l = (q_{2l-1}, q_{2l}), \quad k_l = (q_{2l-1}, q_{2l}),$$

for $l \in [d/2]$, and let ϕ_l be the angle between q_l and k_l, with:

$$\cos\phi_l = \frac{\langle q_l, k_l\rangle}{\|q_l\|_2\|k_l\|_2}\,.$$

Without loss of generality, we make the following assumption:

A3 There exists $\beta_q, \beta_k > 0$ such that $\|q_l^{(t)}\|_2 \geq \beta_q\|q^{(t)}\|_2$ and $\|k_l^{(t)}\|_2 \geq \beta_k\|k^{(t)}\|_2$ for all $l \in [d/2]$ for all $t \geq 0$.

The condition means that all segments makes a nontrivial contribution to the norm. In practice, since LLMs tend to tend to predominantly utilize feature dimensions that rotate slowly (Barbero et al., 2024b), the effective $d/2$ tends to be a small number.

Given the pre-defined set of base rotational angles $\Theta = \{0 \leq \theta_1 \leq \cdots \leq \theta_{d/2}\}$, we reparametrize as $\theta_i = \alpha_i\theta_1$.

H.1 RESULTS

We present the general version of Lemma 3.4 and Theorem 3.5 as follows:

Lemma H.1. *Let $\mathcal{G}$ be the causal mask and* **A1**-**A3** *hold. Suppose for $t \geq 0$, $\|q_i\|_2, \|k_j\|_2 > 0$, and $|\phi_{i,j}^{(t)}| \leq \delta\theta_1$, where $\delta > 0$ and*

$$\left(\sqrt{\frac{1}{\beta_q\beta_k}}\delta\pi + 2(N-1)\alpha_{d/2}\right)\theta \leq 2\pi\,.$$

Then there exists $C_{\max}, C_{\min}, c, c' > 0$ such that

$$C_{\min}e^{-c\sum_{l=1}^{d/2}(i-j)^2\alpha_l^2\theta_1^2} \leq (A_{\mathrm{RoPE}}^{(t)})_{i,j} \leq C_{\max}e^{-c'\sum_{l=1}^{d/2}(i-j)^2\alpha_l^2\theta_1^2}\,.$$

Theorem H.2. *Fix $T > 0$. Under the same conditions as in Lemma H.1 for $t \leq T$, there exists $c > 0$ such that for all $t \leq T$,*

$$(P_{\mathrm{RoPE}}^{(t)})_{i,j} = \Theta\left(\binom{t+i-j}{i-j}e^{-c\sum_{l=1}^{d/2}(i-j)^2\alpha_l^2\theta_1^2}\right)\,.$$

H.2 PROOFS OF LEMMA H.1

We first show the following auxiliary result:

Lemma H.3. *Under* **A3***, it holds that*

$$|\phi_l| \leq \frac{\pi}{2}\sqrt{\frac{1}{\beta_q\beta_k}}|\phi|\,,$$

for all $l \in [d/2]$.

Proof. By definition, since

$$\sum_{l=1}^{d/2}\|q_l\|_2\|k_l\|_2\cos\phi_l = \|q\|_2\|k\|_2\cos\phi\,,$$

then the Cauchy–Schwarz inequality implies that

$$\sum_{l=1}^{d/2}\|q_l\|_2\|k_l\|_2(1-\cos\phi_l) \leq \|q\|_2\|k\|_2(1-\cos\phi)\,. \tag{11}$$

By **A3**, (11) becomes

$$\sum_{l=1}^{d/2}(1-\cos\phi_l) \leq \frac{1}{\beta_q\beta_k}(1-\cos\phi)\,.$$

Given the trigonometric identity $1-\cos 2x = 2\sin^2 x$, we get that

$$\sum_{l=1}^{d/2}\sin^2\left(\frac{\phi_l}{2}\right) \leq \frac{1}{\beta_q\beta_k}\sin^2\left(\frac{\phi}{2}\right)\,. \tag{12}$$

Notice for all $x \in \mathbb{R}$,

$$\sin^2 x \leq x^2\,, \tag{13}$$

and for all $|x| \leq \pi/2$,

$$\frac{4}{\pi^2}x^2 \leq \sin^2 x\,. \tag{14}$$

Apply (13) and (14) to (12), we get that

$$\sum_{i=1}^{d/2}\phi_l^2 \leq \frac{\pi^2}{4\beta_q\beta_k}\phi^2\,.$$

Hence for all $l \in [d/2]$, it follows that

$$|\phi_l| \leq \frac{\pi}{2}\sqrt{\frac{1}{\beta_q\beta_k}}|\phi|\,.$$

□

Denote the angle after rotation to be $\psi_{i,j,l}$. Then it follows that

$$\psi_{i,j,l} = \phi_{i,j,l} - (i-j)\alpha_l\theta_1\,.$$

It follows similarly as in the proof of Lemma 3.4 that

$$C_{\min}\left(1-\delta'^2\theta^2-(i-j)^2\alpha_l^2\theta_1^2\right) \leq \langle (q_i')_l, (k_j')_l\rangle \leq C_{\max}\left(1-\frac{((i-j)^2\alpha_l^2/2-\delta'^2)\theta_1^2}{c}\right)\,,$$

where $\delta' = \frac{\pi}{2}\sqrt{\frac{1}{\beta_q\beta_k}}\delta$, for all $l \in [d/2]$.

Since

$$\langle q_i', k_j'\rangle = \sum_{l=1}^{d/2}\langle (q_i')_l, (k_j')_l\rangle,$$

we conclude the statement.

H.3 Proof of Theorem H.2

The result is a direct corollary of Lemma 3.4 and Theorem 3.5.

I Experiments

Here we provide more details on the numerical experiments presented in Section 4. All models were implemented with PyTorch (Paszke et al., 2019).

Parameterizing the data distribution As defined in Section 4, the input data distribution is modulated by tuning various parameters. In addition to the parameters described in the main text, for the Gaussian mixture with K classes, each class k is defined by a d-dimensional vector μ_k whose components are sampled $i.i.d.$ from a normal distribution with mean zero and variance $1/d$. Then the value of x_i is given by $\frac{\mu_k+\gamma\eta}{\sqrt{1+\gamma^2}}$, where η is drawn from the same distribution as the μ_k's and γ sets the within-class variability. Each class is assigned to one of L labels ($L \leq K$). The contents of the labels are drawn prior to training from the same distribution as the μ_k's.

In Reddy (2024), the author found that different configurations of the data generating process give rise to different learning regimes. To enable better information retrieval ability of the model, we choose the configuration suggested by Reddy (2024) that corresponds to the difficult in-weight learning and easy in-context-learning regime to ensure the information retrieval ability of the model. Specifically, we set $\gamma = 0.75$, $K = 2048$, $L = 32$, and $B = 4$.

Relative PE hyperparameters For the decay mask, we set $m = 0.8$. For RoPE, we set $\theta_i = 10000^{-2(i-1)/d}$, as in Su et al. (2023).

Compute We trained all of our models on a Tesla V100 GPU.

Training details In all experiments, we used the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 10^{-3}, a weight decay of 10^{-6}, a batch size of 128, and trained for $100,000$ iterations.

J Additional Experimental Results

J.1 The role of training data on positional bias

In this section, we present additional experimental results building on the experiment described in Section 4, but focusing on cases where positional bias in the training sequences is introduced at specified positions. Specifically, we consider four types of training sequences where x_{query} is assigned the class of 1) x_1 (the first position) or x_n (the last position) with equal probability, 2) x_1 , 3) $x_{n/2}$ (the middle position), or 4) x_n. The corresponding average results using a 2-layer network over five runs are shown in Figure 3, Figure 4, Figure 5, and Figure 6, respectively.

Observe that, compared with no mask, the causal mask without PEs indeed introduces a sense of position across all cases. Specifically, it enables the model to learn a positional bias favoring the beginning of the sequence, as earlier tokens tend to receive more attention through the mechanism of iterative attention. This is particular notable in Fig. 3, where a causal mask without PE only enables the model to learn a position bias at the beginning of the sequence but not at the end. In contrast, both sin PE and RoPE allow the model to effectively capture different positional biases regardless of their location in the training sequences.

The role of data in creating positional bias It is worth noting that in Figure 3, we observe the "lost-in-the-middle" phenomenon (Liu et al., 2024), where information retrieval accuracy follows a U-shape relative to the position of the answer, with performance at the beginning of the sequence slightly better than at the end. We note that the "lost-in-the-middle" phenomenon only emerges when the training sequences are biased toward both the beginning and the end. This suggests that specific

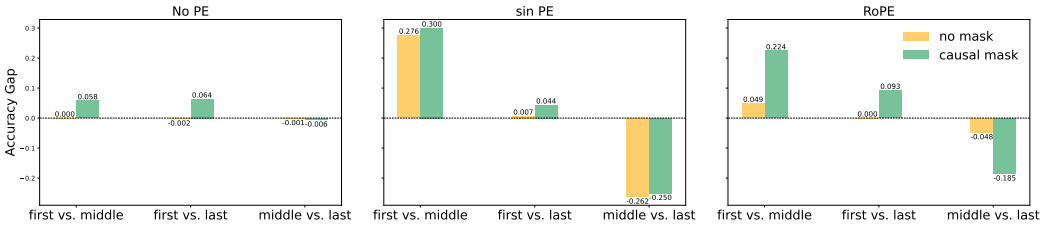

Figure 3: Position bias when trained on data biased toward both the **first** and **last** positions.

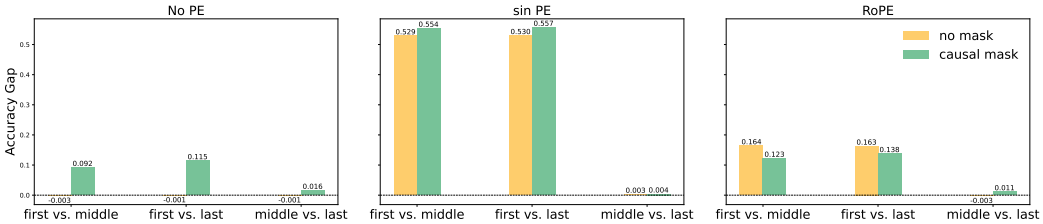

Figure 4: Position bias when trained on data biased toward the **first** position.

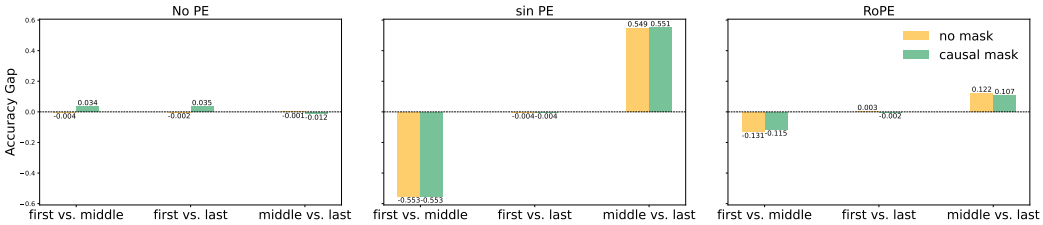

Figure 5: Position bias when trained on data biased toward the **middle** position.

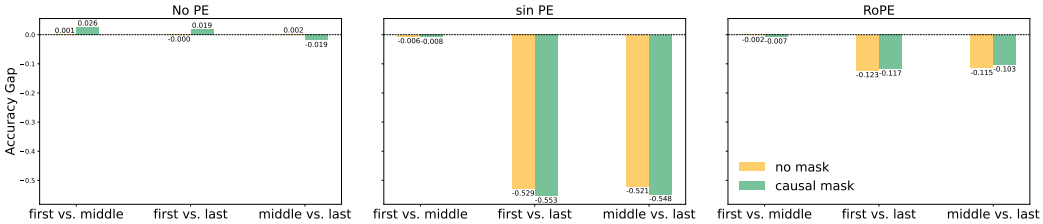

Figure 6: Position bias when trained on data biased toward the **last** position.

types of positional bias in the training data play a crucial role in shaping how the model learns to process and prioritize positions within a sequence.

As the structure of positional bias in natural language remains unclear, this observation raises the following question:

> *Does positional bias in natural language sequences shape the "lost-in-the-middle" phenomenon in a similar way to what we observe in this simplified case?*

This question connects to broader inquiries about the parallels between artificial and human attention. In neuroscience, the primacy-recency effect highlights that human attention often gravitates toward the beginning and end of sequences (Glanzer and Cunitz, 1966; Li et al., 2024), a phenomenon that may have influenced the structure of human languages, where critical information is frequently placed in these positions (Halliday, 2004). As demonstrated in Figure 3, when such patterns are present in training data, attention-based architectures seem to develop analogous biases (Hollenstein et al., 2021), aligning with natural language characteristics for improved performance. This raises deeper, perhaps philosophical questions: To what extent are these biases intrinsic to effective sequential processing? How closely should neural networks emulate human cognitive patterns? Investigating these connections can deepen our understanding of both human and artificial intelligence while guiding the design of more effective machine learning models.

J.2 Attention sinks

Despite our use of a simplified experimental setup in this work, we observe the emergence of key phenomena documented in more complex settings. In addition to the "lost-in-the-middle" phenomenon discussed in Appendix J.1, in this section, we report the formation of attention sinks in our setting.

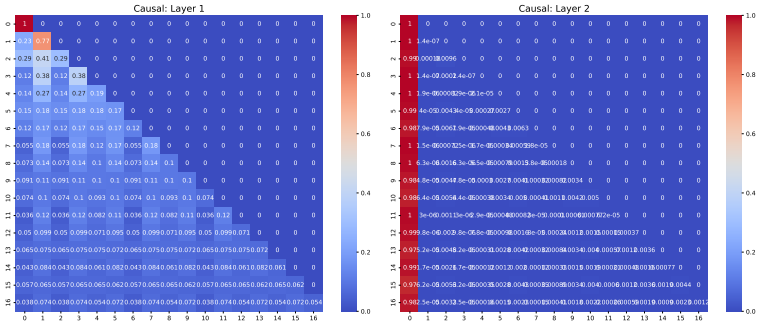

Figure 7: Example of the emergence of attention sinks in our experimental setting. In particular, the sequences used for training and inference are all **free of position bias**.

Figure 7 shows an example of the attention maps of a two-layer self-attention networks under the causal mask without PEs, where the sequences used for training and inference are all free of position bias. We observe the similar phenomenon of attention sinks as reported in Xiao et al. (2024).

More quantitatively, following Gu et al. (2025), we calculate their metric for measuring the emergence of attention sinks, over $10,000$ sequences free of position bias. Specifically, denote the adjacency matrix of the mask $\mathcal{G}$ to be M. Then the metric for attention sink at token j is calculated as

$$\text{Attention Sink}_j = \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\sum_{i=1}^{N} M_{ij}}\sum_{i=1}^{N}\mathbf{1}\{A_{ij}^{(t)} > \tau\}\,.$$

The threshold τ we choose is 0.2. The results for the causal mask, the sliding-window masks (with $w = 5, 9, 13$), and the prefix masks (with $K = 2, 4, 6$) are shown in Figures 8, 9, and 10, respectively. In particular, we make the following observations:

1. Attention sinks emerge on the absolute first token under the causal mask.
2. Attention sinks tend to emerge on the absolute first token when the window size w is larger, under the sliding-window mask.
3. Attention sinks emerge on the K prefix tokens, not just on the first token alone, under the prefix mask.

All of these phenomena have been observed in real-world LLMs in Gu et al. (2025). This alignment between our controlled setup and real-world observations affirms the validity of our abstraction, indicating that we have captured the key mechanisms underlying position bias while facilitating a systematic analysis.

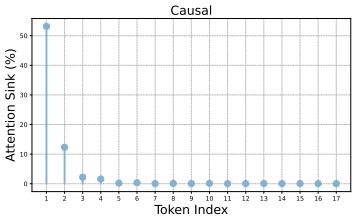

Figure 8: Attention sinks emerge on the first token under the causal attention mask.

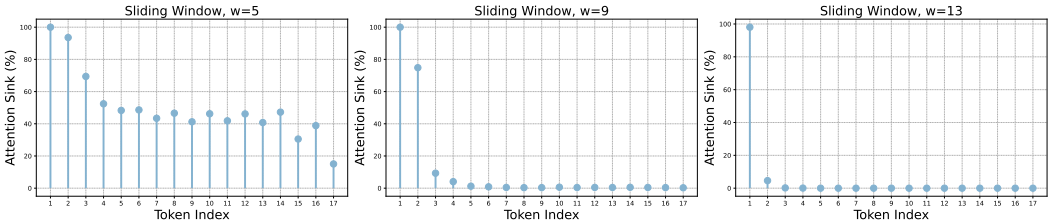

Figure 9: Attention sinks tend to emerge on the absolute first token when the context window size w is larger, under the sliding-window mask.

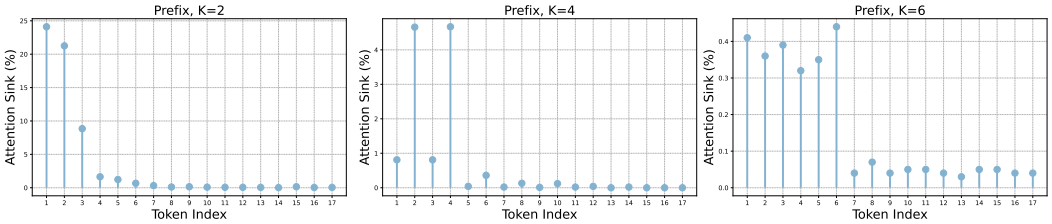

Figure 10: Attention sinks emerge on the K prefix tokens, not just on the first token alone, under the prefix mask.